CAPSTONE PROJECT-4

Book Recommendation System

By Jaya Vishwakarma

Problem Statement

During the last few decades, with the rise of YouTube, Amazon, Netflix, and many other such web services, recommender systems have become much more important in our lives in terms of providing highly personalized and relevant content. The main objective is to create a recommendation system to recommend relevant books to users based on popularity and user interests.

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries). Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. The main objective is to create a book recommendation system for users.

Content

- ✓ Problem Statement & Data Description
- ✓ Analysis of different datasets.
  1. Books
  2. Ratings
  3. Users
- ✓ Data Cleaning
- ✓ Outlier treatment
- ✓ Imputing missing values
- ✓ Different Recommendation Model
- ✓ Challenges
- ✓ Conclusion

Data Summary

Books Data

ISBN – We have a unique ISBN number for all the books.
Book Title – Book title correspond to the ISBN number.
Book Author – Name of book author
Year of Publication – In which year book published
Publisher - Publishing company/house name
URL Links (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large.

Users Data

User ID – A unique user id of all the users
Location – City, state and country of the user
Age – Age of the user

Ratings

Book Rating – Rating provide by the user for a particular book between 0-10
User ID & ISBN – Basically this to map book rating with other 2 data

Data Info

```
# Books data info
books.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271360 entries, 0 to 271359
Data columns (total 5 columns):
 #   Column               Non-Null Count   Dtype 
---  ------               --------------   ----- 
 0   ISBN                 271360 non-null  object
 1   Book-Title           271360 non-null  object
 2   Book-Author          271359 non-null  object
 3   Year-Of-Publication  271360 non-null  object
 4   Publisher            271358 non-null  object
dtypes: object(5)
memory usage: 10.4+ MB
```

```
# Rating data info
ratings.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1149780 entries, 0 to 1149779
Data columns (total 3 columns):
 #   Column       Non-Null Count    Dtype 
---  ------       --------------    ----- 
 0   User-ID      1149780 non-null  int64 
 1   ISBN         1149780 non-null  object
 2   Book-Rating  1149780 non-null  int64 
dtypes: int64(2), object(1)
memory usage: 26.3+ MB
```

```
# User data info
users.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 278858 entries, 0 to 278857
Data columns (total 3 columns):
 #   Column    Non-Null Count   Dtype  
---  ------    --------------   -----  
 0   User-ID   278858 non-null  int64  
 1   Location  278858 non-null  object 
 2   Age       168096 non-null  float64
dtypes: float64(1), int64(1), object(1)
memory usage: 6.4+ MB
```

The Book-Crossing dataset comprises 3 files.

Books – ISBN, Title, Author, Publisher & year

Ratings – User ID, ISBN, Rating

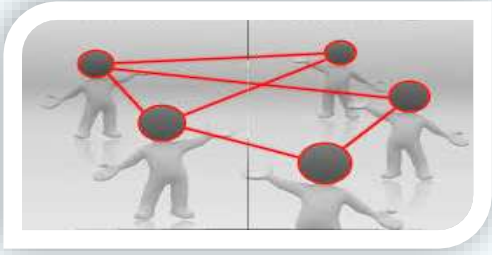
Users - User ID, Location, Age

Data Pipeline

Data Understanding: Summarize the data, such as data volume and total number of variables in the data. Understand the problems with the data, such as missing values, inaccuracies, and outliers.

Data Processing: Checked duplicate, null values and detection of outliers. Replaced null values and treated outliers. There are some incorrect entries in data which was replaced with correct values.

EDA: It's a critical process of performing the initial investigation on the data. Throughout this process, we have observed certain trends. We also drawn certain conclusions from the dataset that will be useful for further processing.



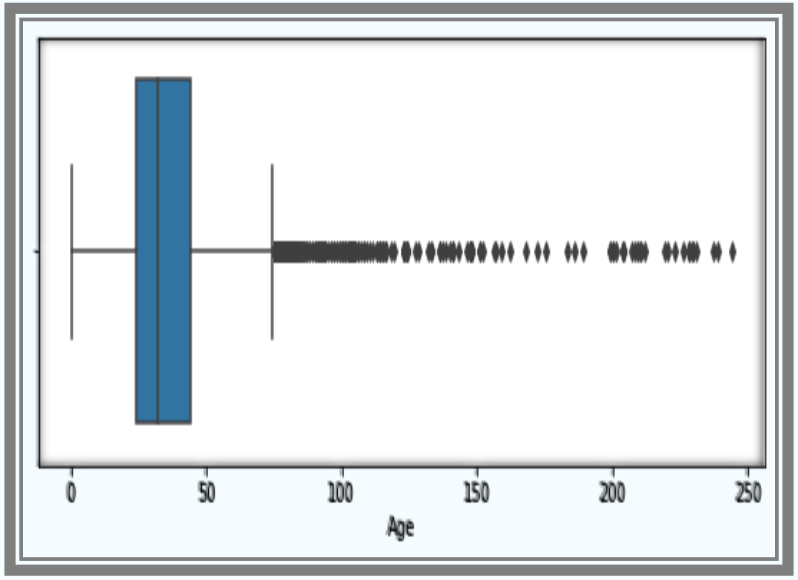
Recommendations

- First, we have **popularity based recommendation** where we have taken out top 20 books, author and publisher on which more than 250+ experienced users have rated. By saying experienced users, we are referring to those users who have rated more than 250 books. They can be reviewers, proof-reader, literary critic etc.

- Second, we have **collaborative filtering based recommendation** which is to recommend similar books with respect to another book. This technique can filter out items that users like on the basis of the ratings by similar users.

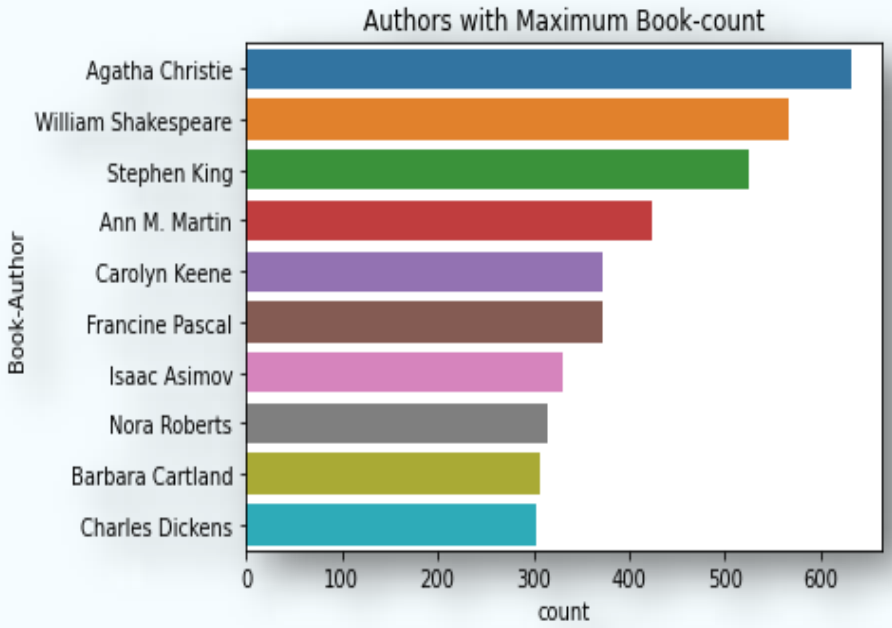
- At last, we have **model based collaborative filtering recommender**. This is to predict user's preference for a set of items based on the past experience. Model based approach involves building machine learning algorithms to predict user's ratings. They involve dimensionality reduction methods that reduce high dimensional matrix containing abundant number of missing values with a much smaller matrix in lower-dimensional space. The goal of this section is to compare **SVD and NMF algorithms**, try different configurations of parameters and explore obtained results.

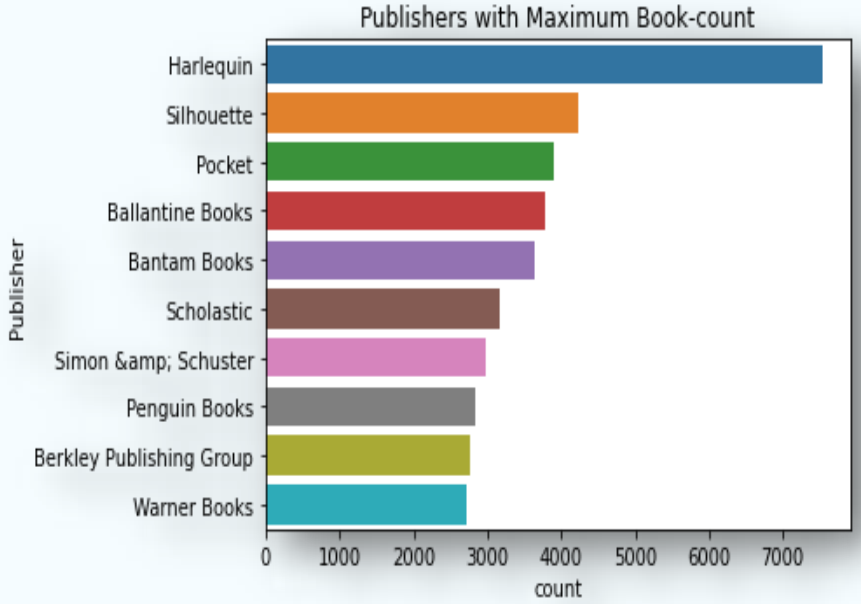
Outliers Detection



The main reason for this detection is that the outliers can **cause serious issues in statistical analysis**. Hence we have checked this before starting analysing our data. So that we can proceed further without having measurement errors, data entry or processing errors.
In our user data, we have **outliers** in **Age** column. The Age range given here is from 0 to 250. Which was treated by assigning mean values.

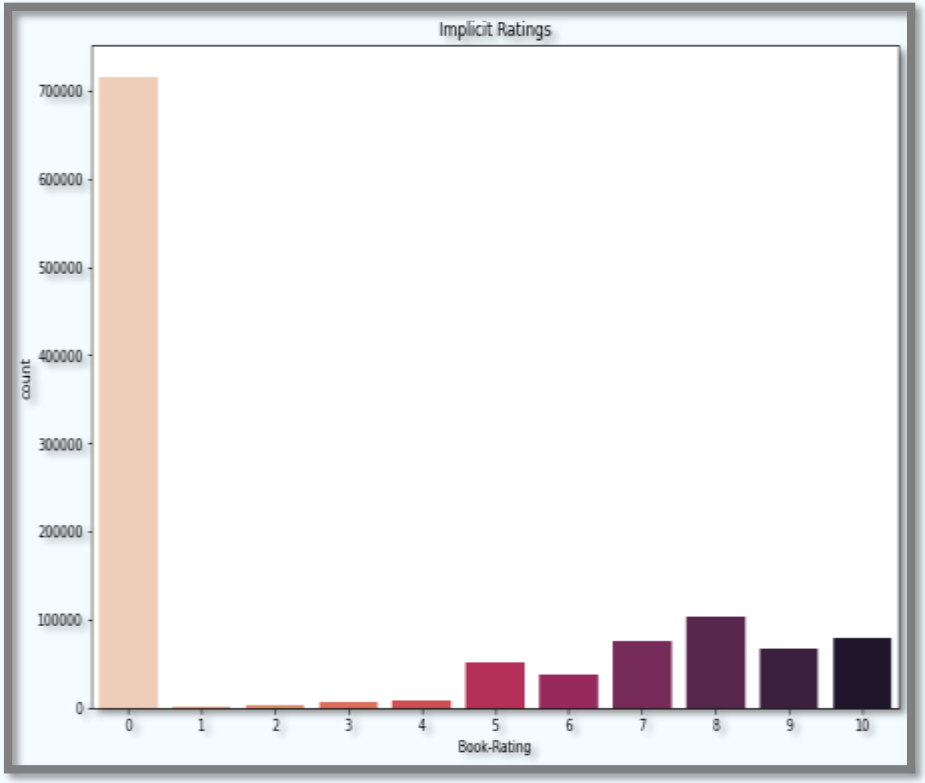
EDA: Books data

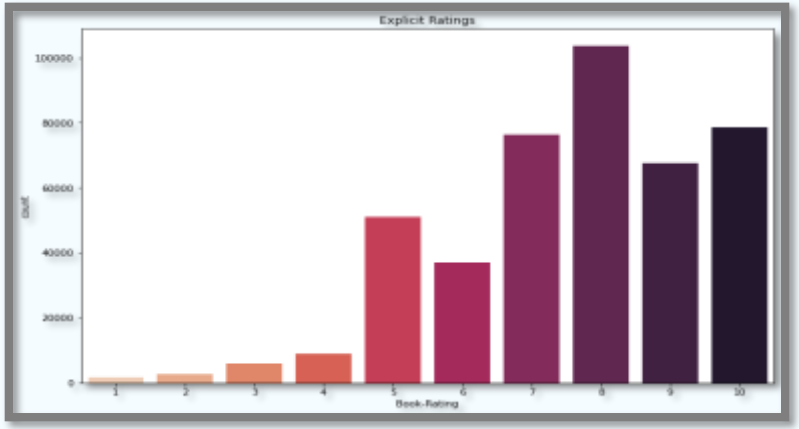




Here's we have count of Author & Publisher with maximum books.
Agatha Christie wrote maximum books followed by William Shakespeare and Stephen King.
Hariequin enterprises published highest number of books.

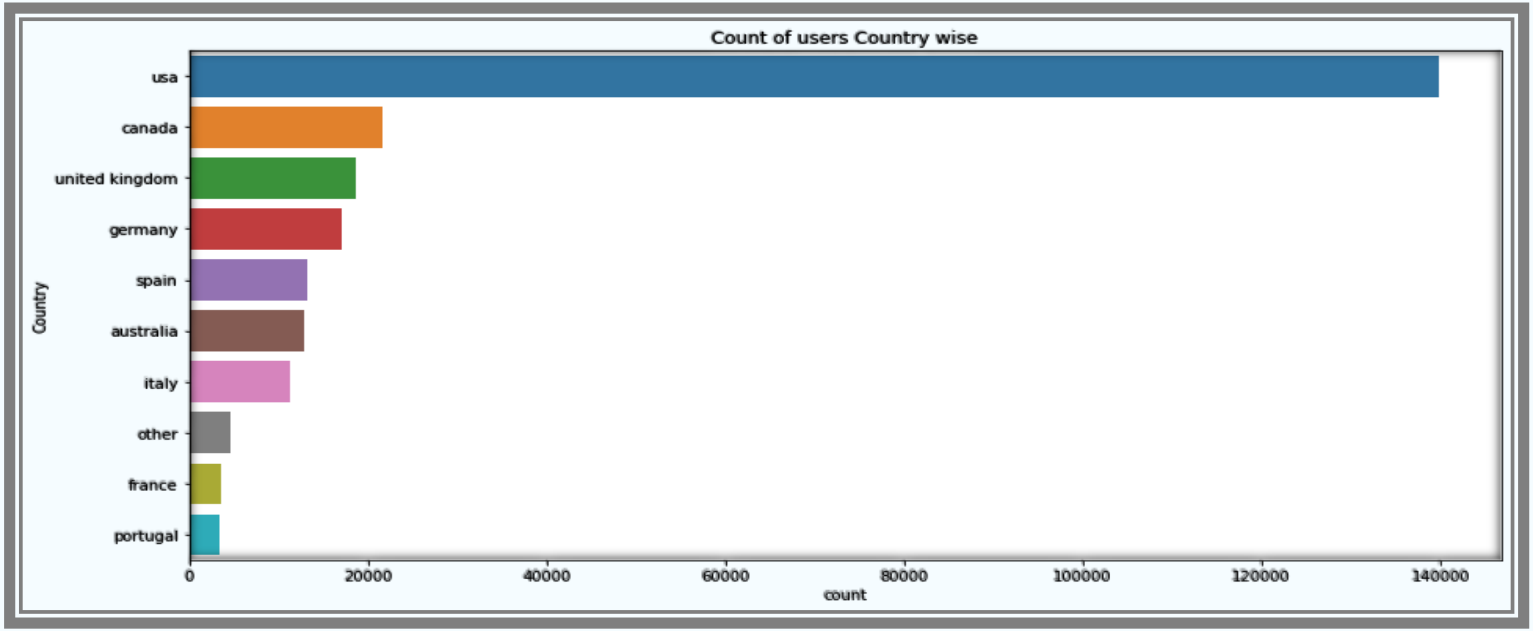
EDA: Ratings data





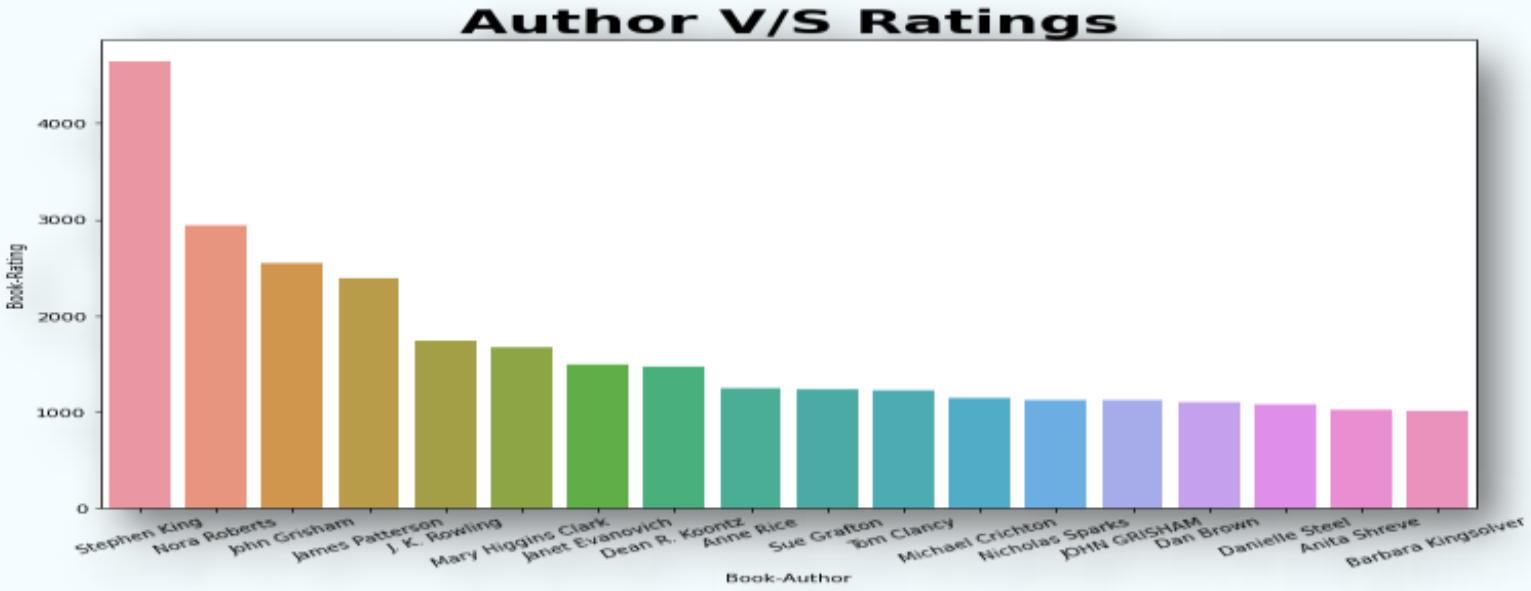
We have implicit rating and explicit rating. Higher ratings are more common amongst users and rating 8 has been rated highest number of times. Very few have rated below 5.

EDA: Users data



Maximum number of users are from **USA** followed by **Canada** and **United Kingdom**.

Author to Receive Maximum No. of Ratings

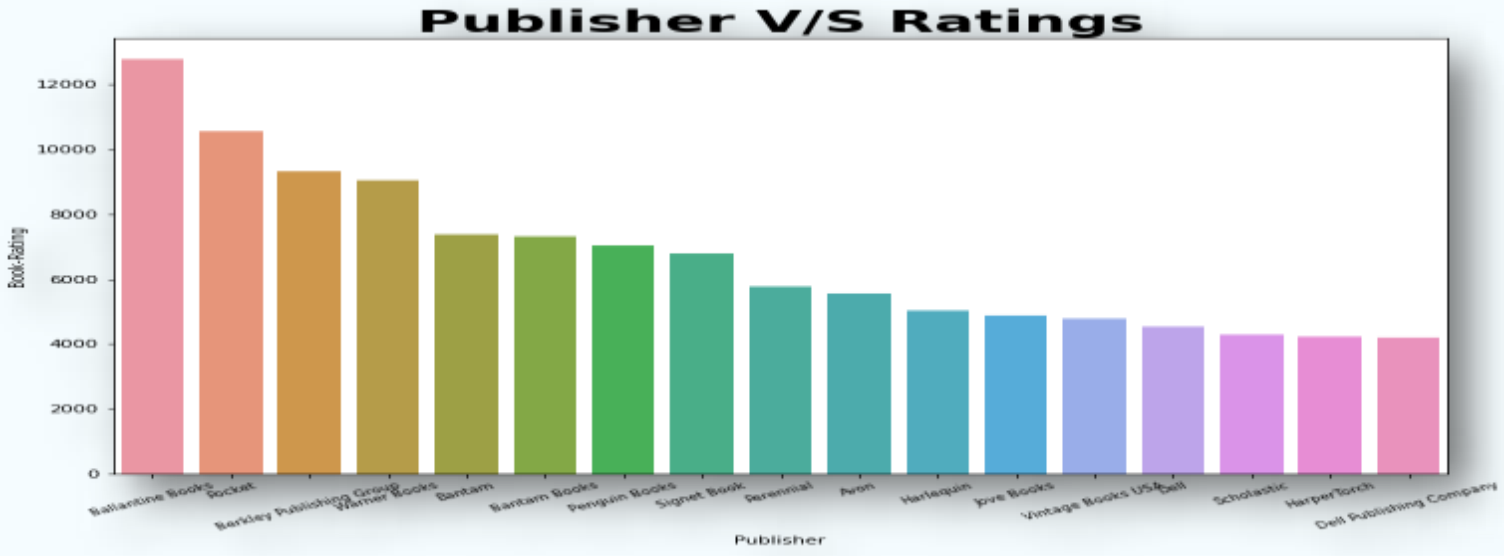


Stephen king is the author to receive maximum ratings by the users. Then we have Nora Roberts, John Grisham, James Patterson and J.K. Rowling.

Books to Receive Maximum No. of Ratings

Top-10 most rated books were essentially **novels**. Books like **The Lovely Bone** and **Wild Animus** were very well perceived.

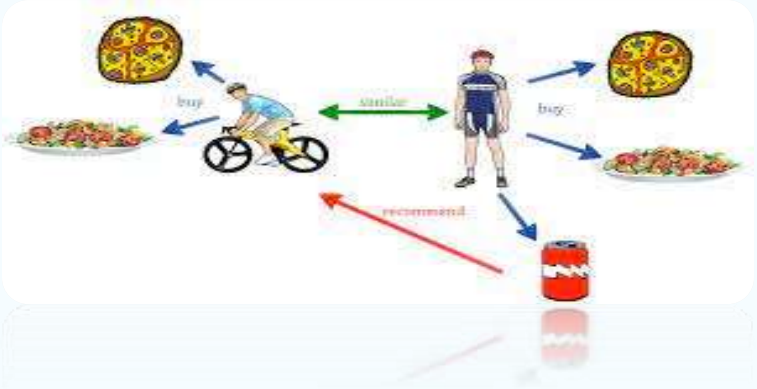
Publisher to Receive Maximum No. of Ratings



Here we have top-10 most rated Publishers. Ballantine Books received maximum attention by the users followed by Rocket and Barkley Publishing Group.

Recommendation System

Build function and models: popularity based recommendation, collaborative filtering based recommendation and
model based collaborative filtering recommender.



Popularity Based: 10 Most Popular Books

	Book-Title	avg_rating	ISBN	Book-Author	Year-Of-Publication	Publisher
0	Harry Potter and the Prisoner of Azkaban (Book 3)	9.043321	0439136350	J. K. Rowling	1999.0	Scholastic
277	To Kill a Mockingbird	8.977528	0446310786	Harper Lee	1988.0	Little Brown & Company
544	Harry Potter and the Sorcerer's Stone (Harry P...	8.936508	059035342X	J. K. Rowling	1999.0	Arthur A. Levine Books
859	Harry Potter and the Chamber of Secrets (Book 2)	8.840491	0439064864	J. K. Rowling	1999.0	Scholastic
1185	Tuesdays with Morrie: An Old Man, a Young Man,...	8.588000	0385484518	MITCH ALBOM	1997.0	Doubleday
1435	The Secret Life of Bees	8.477833	0142001740	Sue Monk Kidd	2003.0	Penguin Books
1841	The Da Vinci Code	8.439271	0385504209	Dan Brown	2003.0	Doubleday
2335	The Lovely Bones: A Novel	8.185290	0316666343	Alice Sebold	2002.0	Little, Brown
3042	The Red Tent (Bestselling Backlist)	8.182768	0312195516	Anita Diamant	1998.0	Picador USA
3425	Where the Heart Is (Oprah's Book Club (Paperba...	8.142373	0446672211	Billie Letts	1998.0	Warner Books

We have sample of **top 10 books**. More than **250 users have rated** on the these top books. Maximum rating is 10 and minimum is 7.

10 Most Popular Authors

	Book-Title	avg_rating_author	ISBN	Book-Author	Year-Of-Publication	Publisher
0	Harry Potter and the Chamber of Secrets (Book 2)	8.970218	0439064864	J. K. Rowling	1999.0	Scholastic
1746	Pigs in Heaven	8.195437	0060168013	Barbara Kingsolver	1993.0	Harpercollins
2754	The Lovely Bones: A Novel	8.171336	0316666343	Alice Sebold	2002.0	Little, Brown
3682	Angels & Demons	8.116848	0671027360	Dan Brown	2001.0	Pocket Star
4786	Full House	7.944966	1559277785	Janet Evanovich	2002.0	Audio Renaissance
6276	The Dark Half	7.815046	0451167317	Stephen King	1994.0	Signet Book
10915	The Rescue	7.739169	0446610399	Nicholas Sparks	2001.0	Warner Books
12046	B Is for Burglar (Kinsey Millhone Mysteries (P...	7.722267	0553280341	Sue Grafton	1986.0	Bantam
13281	The Beach House	7.697947	0446612545	James Patterson	2003.0	Warner Books
15668	The Street Lawyer	7.640179	0440225701	JOHN GRISHAM	1999.0	Dell

Here, we have sample data of **top 10 Author**. More than **900 users have rated** on the these authors books.

10 Most Popular Publisher

	Book-Title	Book-Rating	ISBN	Book-Author	Year-Of-Publication	Publisher
0	Tell Me This Isn't Happening	5	0439095026	Robynn Clairday	1999.0	Scholastic
4284	My First Cousin Once Removed: Money, Madness, ...	4	0060930365	Sarah Payne Stuart	1999.0	Perennial
10059	McDonald's: Behind the Arches	9	0553347594	John F. Love	1995.0	Bantam
17442	Snow Angels	8	0140250964	Stewart O'Nan	1995.0	Penguin Books
24462	A Judgement in Stone	8	0375704965	Ruth Rendell	2000.0	Vintage Books USA
29232	Rebecca	10	0380778556	Daphne Du Maurier	1994.0	Avon
34797	Airframe	9	0345402871	Michael Crichton	1997.0	Ballantine Books
47578	Fast Women	8	0312252617	Jennifer Crusie	2001.0	St. Martin's Press
51029	The Pillars of the Earth	3	0451166892	Ken Follett	1996.0	Signet Book
57804	This Year It Will Be Different: And Other Stories	8	0440223571	Maeve Binchy	1997.0	Dell

Here, we have sample data of **top 10 Publishers**. More than **3400 users** have given **maximum ratings** on the books published by them.

Collaborative filtering based recommendation

```
recommend('Harry Potter and the Chamber of Secrets (Book 2)')

[['Harry Potter and the Prisoner of Azkaban (Book 3)',
  'J. K. Rowling',
  'Scholastic'],
 ['Harry Potter and the Goblet of Fire (Book 4)',
  'J. K. Rowling',
  'Scholastic'],
 ["Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))",
  'J. K. Rowling',
  'Arthur A. Levine Books'],
 ["Harry Potter and the Sorcerer's Stone (Book 1)",
  'J. K. Rowling',
  'Scholastic'],
 ['Harry Potter and the Order of the Phoenix (Book 5)',
  'J. K. Rowling',
  'Scholastic']]
```

Here's the result after applying collaborative filtering. And we are getting recommendation of similar books.

It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user.
We have applied both the types.
User-based, which measures the similarity between target users and other users.
Item-based, which measures the similarity between the items that target users rate or interact with and other item.

Model Based Collaborative Filtering Recommender

This recommendation system tries to find similarities between users or between items based on recorded user-item preferences or ratings.

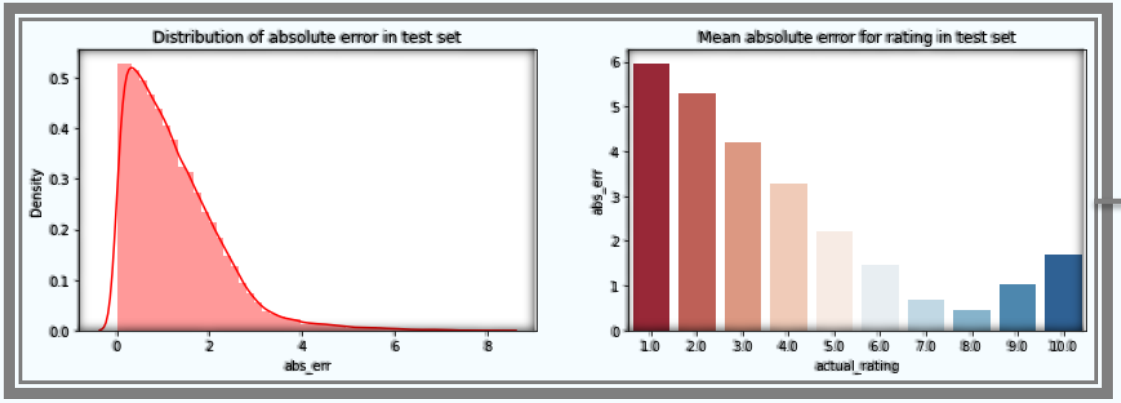
Model based approach involves building machine learning algorithms to predict user ratings. They involve dimensionality reduction methods that reduce high dimensional matrix containing abundant number of missing values with a much smaller matrix in lower-dimensional space.

The goal of this section is to compare SVD and NMF algorithms, try different configurations of parameters and explore obtained results.

Test set: predicted top rated books

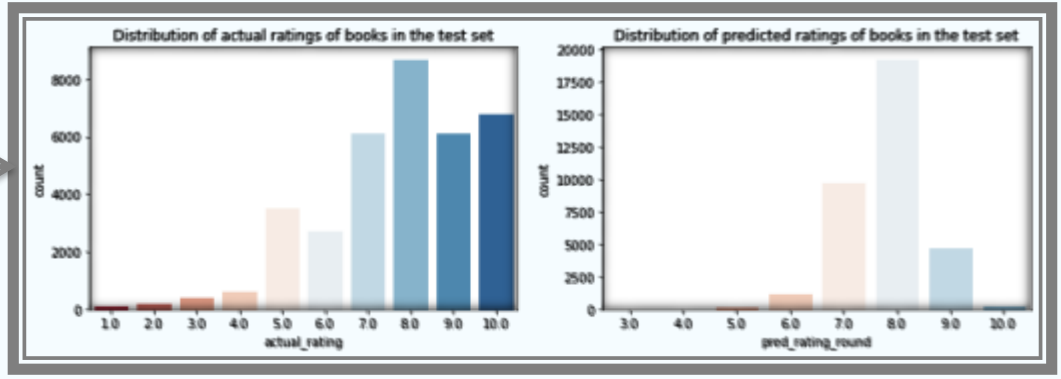
	user_id	isbn	book_rating	book_title	pred_rating
124976	193458	1853260622	5	War and Peace (Wordsworth Classics)	8.187135
124962	193458	0671880314	9	Schindler's List	8.183804
124920	193458	006447108X	9	The Last Battle	8.036723
124965	193458	0767904133	8	Close to Shore: A True Story of Terror in an A...	7.885211
124975	193458	1853260169	10	Sense and Sensibility (Wordsworth Classics)	7.813416

Model Evaluation: Error & Comparison



Mean absolute error

Actual V/s Predicted



Challenges

- Handling of sparsity was a major challenge since, the user interactions were not present for the majority of the books.
- Understanding the metric for evaluation was a challenge as well.
- Since the data consisted of text data, data cleaning was a major challenge in features like Location etc. Also we had some incorrected data for 2-3 books.
- Decision making on missing value imputations and outlier treatment was quite challenging as well.

Conclusion

Top-10 most rated books were essentially novels. Books like **“The Lovely Bone”** and **“Wild Animus”** were very well perceived.

Maximum number of **users** are from **USA** followed by Canada and United Kingdom.

If we look at the ratings distribution, most of the books have high ratings with **maximum books** being **rated 8**. Ratings below 5 are few in number.

Stephen king is the author to receive **maximum ratings** by the users. Then we have Nora Roberts, John Grisham, James Patterson and J.K. Rowling.

For modelling, it was observed that for model based collaborative filtering **SVD techn ique** worked way better than NMF with **lower** Mean Absolute **Error** (MAE).

THANK YOU!